How reliably can we infer diversity-dependent diversification from phylogenies?

Etienne, Rampal S.; Pigot, Alex L.; Phillimore, Albert B.

Published in:
Methods in ecology and evolution

DOI:
10.1111/2041-210X.12565

Publication date:
2016

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7,** 1092–1099 doi: 10.1111/2041-210X.12565

How reliably can we infer diversity-dependent diversification from phylogenies?

Rampal S. Etienne[1]*, Alex L. Pigot[1] and Albert B. Phillimore[2]

[1]*Groningen Institute for Evolutionary Life Sciences, University of Groningen, PO Box 11103, Groningen 9700 CC, The Netherlands; and* [2]*Institute of Evolutionary Biology, University of Edinburgh, Edinburgh EH9 3JT, UK*

Summary

1. Slowdowns in lineage accumulation in phylogenies suggest that speciation rates decline as diversity increases. Likelihood methods have been developed to detect such diversity dependence. However, a thorough test of whether such approaches correctly infer diversity dependence is lacking.

2. Here, we simulate phylogenetic branching under linear negative diversity-dependent and diversity-independent models and estimate from the simulated phylogenies the maximum-likelihood parameters for three different conditionings – on survival of the birth–death process given the crown age, on tree size (N) and on tree size given the crown age. We report the accuracy of recovering the simulation parameters and the reliability of the model selection based on the χ^2 likelihood ratio test.

3. Parameter estimate accuracy: Conditioning on survival given the crown age yields a severe bias of the carrying capacity K towards N and an upward bias of the speciation rate, particularly in clades where diversity-dependent feedbacks are still weak ($N \ll K$). Conditioning on N yields an overestimate of K and an underestimate of speciation rate, particularly when saturation has been reached. Dual conditioning yields relatively unbiased parameter estimates on average, but the deviation from the true value for any single estimate may be large.

4. Model selection reliability: The frequency of incorrectly rejecting a diversity-independent model when the simulation was diversity-independent (type I error) differs substantially from the significance level α used in the likelihood ratio test, rendering the likelihood ratio test inappropriate. The frequency of correctly rejecting the diversity-independent model when the simulation was diversity-dependent (power) is larger when the clade is closer to equilibrium and for conditioning on crown age.

5. We conclude that conditioning on crown age has the best statistical properties overall, but caution that parameter estimates may be biased. To assess parameter uncertainty in future studies of diversity dependence on real data, we recommend parametric bootstrapping, examination of the likelihood surface and comparison of estimates across the types of conditioning. To assess model selection reliability, we discourage the use of the χ^2 likelihood ratio test or AIC (which are equivalent in this case), but recommend a likelihood ratio test based on parametric bootstrap. We illustrate this method for the diversification of *Dendroica* warblers.

Key-words: Birth–death model, conditioning, diversity dependence, extinction, parametric bootstrap, simulations

Introduction

Over the last few decades, the use of molecular phylogenies to understand the dynamics of lineage diversification has grown rapidly (Rabosky 2009; Morlon 2014). A common observation is that clades appear to undergo an initial burst of diversification followed by a slowdown towards the present (Weir 2006; Phillimore & Price 2008). This pattern has attracted considerable interest and is often interpreted as evidence that as a clade radiates and niches become filled, clade diversity feeds back negatively on the opportunities for further speciation (Phillimore & Price 2008; Rabosky & Lovette 2008a; Rundell & Price 2009; Rabosky 2013), although several alternative explanations exist (Pigot *et al.* 2010; Etienne & Rosindell 2012; Moen & Morlon 2014; Harmon & Harrison 2015).

Until recently, many tests for diversification slowdowns employed the gamma statistic, which was designed to test whether the temporal distribution of branching events on a reconstructed phylogeny departed from the expectation under a pure-birth process (Pybus & Harvey 2000). More recently, models have been developed to explicitly test whether rates of diversification depend on the number of extant lineages at any given point in time. While the first of these 'diversity-dependent' models ignored extinction (Rabosky & Lovette 2008a; Bokma 2009), new approaches have been developed that allow for non-zero extinction and thus account for extinct lineages, which have left no extant descendants but may still have influenced the dynamics of diversification in the past (Etienne & Haegeman 2012; Etienne *et al.* 2012). This modelling

*Correspondence author. E-mail: r.s.etienne@rug.nl

framework reconciles two widespread observations of diversity dynamics: that many clades undergo temporal declines in diversification and that extinction is a pervasive feature of the fossil record.

Diversity-dependent models are generalizations of the constant-rate (diversity-independent) birth–death process allowing for a decline in the per-lineage speciation rate as the number of lineages accumulates (Rabosky & Lovette 2008a). Increases in extinction rate with diversity are also possible, but these generally find little support from molecular phylogenetic data (Rabosky & Lovette 2008b). In addition to the initial speciation rate and the extinction rate, this model requires an additional parameter K, representing the equilibrium species richness or 'carrying capacity' of the clade. Formally, K corresponds to the equilibrium value of N (the number of extant species) where the rate of speciation equals the rate of extinction. Alternatively, we can introduce a parameter K' that denotes the maximum (gamma) diversity of species that could exist in the absence of extinction. If all species are sympatric, then a possible interpretation of K' is the total number of available niches (although in this context the niche concept is often not clearly defined, McInerny & Etienne 2012a,b,c). In contrast, if all species are allopatric, then K' may be interpreted as the number of islands or regions that could generate allospecies. K and K' are related to one another, depending on the functional form of the diversity dependence, and below we give an example for linear negative diversity dependence of the speciation rate.

Given the observed branching times in an empirical phylogeny, the likelihood of the diversity-dependent speciation model and simpler constant-rate pure-birth or birth–death models can be calculated (Etienne *et al*. 2012) and model comparison can proceed via likelihood ratio tests or AIC. In addition to testing for evidence of diversity dependence, these models also return parameter estimates that potentially offer further insights into the dynamics of a clade's diversification, including the initial speciation rate, the total number of available niches (K'), carrying capacity (K) and whether the number of species in the clade has reached equilibrium ($N/K = 1$) or is still in an ascending phase (Etienne *et al*. 2012; Jønsson *et al*. 2012; Pyron & Wiens 2013; Valente, Phillimore & Etienne 2015). While this represents an exciting possibility, it is important that we understand the degree to which the parameters estimated from these models are reliable and whether model comparison leads to the (most) correct model being inferred.

Previous work exploring the performance of diversity-dependent models suggested that under certain conditions – when clades are young and/or governed by high rates of extinction – K is biased downwards towards the observed tree size (N) and the initial speciation rate (λ_0) is biased upwards (Etienne *et al*. 2012). However, the severity of these potential biases and the conditions under which they arise have yet to be thoroughly explored. Here, we examine this question using simulations across a broad combination of diversification scenarios. We assess the statistical performance of this diversity-dependent model in terms of (i) its precision and bias in recovering the simulation parameters, with a particular focus on K, and (ii) the power to detect diversity dependence.

Because we always analyse extant phylogenies, the likelihood should be conditioned on survival of the birth–death process until the present (Nee, May & Harvey 1994; Stadler 2013). While the standard approach is to additionally condition on the observed crown age, it is also possible to condition on the tree size, N (Stadler 2013), or on both crown age and N. These different forms of conditioning incorporate different amounts of the information contained in the phylogeny into the likelihood, and thus, we also examine how the choice of conditioning may influence statistical performance.

We find that when conditioning is on crown age or N, parameter estimates (particularly for the clade-level carrying capacity K) are often biased. In contrast, simultaneous conditioning on crown age and N ameliorates the biases in parameter estimates but at the expense of greatly reduced precision. Furthermore, we find that standard likelihood ratio tests, or equivalently, AIC-based model comparisons, yield incorrect type I errors. We discuss the implications of these results for inferences regarding diversity-dependent species diversification and provide guidelines for future research. In particular, we recommend the use of parametric bootstrapping for a proper likelihood ratio test, and we illustrate this using the diversification of *Dendroica* (recently moved to the genus *Setophaga*) as a case study (Rabosky & Lovette 2008a; Etienne *et al*. 2012).

Methods

We assumed a birth–death model of diversification with the following rates of speciation and extinction as a function of diversity N:

$$\lambda(N) = \max(0, \lambda_0 - (\lambda_0 - \mu)N/K)$$
$$\mu(N) = \mu$$

where λ_0 is the initial speciation rate. This dependence on N ensures that speciation and extinction rates equal each other at the 'carrying capacity' $\lambda(K) = \mu$. This model is mathematically equivalent to a model with

$$\lambda(N) = \max(0, \lambda_0(1 - N/K'))$$
$$\mu(N) = \mu$$

under the substitution $K' = \lambda_0 K/(\lambda_0 - \mu_0)$. We note that viewing K as the carrying capacity is just one biological interpretation of a mathematical parameter. Furthermore, we simply assume a linear negative diversity dependence, which is arguably the simplest way to incorporate diversity dependence. To assess model performance, we simulated clades under known combinations of parameter values, hereafter for brevity termed 'diversification scenarios'. These were as follows: initial speciation rate λ_0 (0·5, 0·8), extinction rate μ (0, 0·1, 0·2, 0·4), carrying capacity K (40, 80, ∞) and crown age (5, 10, 15). When $K = \infty$, this corresponds to a constant-rate pure-birth ($\mu = 0$) or birth–death ($\mu > 0$) model. When $K = 40$ or $K = 80$, this corresponds to a diversity-dependent scenario with zero or non-zero constant extinction. We varied crown age to produce a range of diversification scenarios from young clades that are far from equilibrium to old clades that have reached equilibrium. We chose these diversification scenarios because they represent cases for which it is reasonable to apply diversity dependence to the whole tree. We argue that for older and larger clades, rate shifts (Alfaro *et al*. 2009) and decoupling of diversity-dependent dynamics (Etienne & Haegeman 2012) will make it highly unlikely that the clade has been governed by a single diversity-dependent process.

For each combination of parameter values, we used the R package DDD, v3.2 (Etienne *et al.* 2012), to simulate 1000 trees and estimate maximum-likelihood parameters for each tree under the (i) constant-rate birth–death model – estimating λ_0 and μ while fixing $K = \infty$ (two parameters), (ii) diversity-dependent model with extinction – estimating λ_0, μ and K (three parameters). We used two starting values for each optimization of the diversity-dependent model to reduce the risk of convergence on a local rather than global optimum and checked manually for several simulations whether higher optima existed, which we found not to be the case. To assess how the form of conditioning influences model performance, we fitted each of these models using three different conditionings on: (i) crown age, (ii) tree size at the present and (iii) crown age and tree size. Mathematically, this means that we divide the likelihood that is conditioned only on crown age (but not survival) by the probability of both crown lineages surviving for conditioning (i) and, for conditioning (iii), by the probability of both crown lineages surviving and yielding precisely N species at the present. For conditioning (ii), we assume a uniform prior on the stem age and integrate the likelihood for conditioning (iii) across all stem ages from $-\infty$ to the observed crown age. This is identical to what has been commonly done for the diversity-independent model when conditioning on tree size (Stadler 2013). In practice, this involves either numerically integrating backwards from the present or computation of the inverse of the large transition matrix, both of which are computationally costly.

For diversification scenarios where diversity dependence was operating ($K = 40$ or $K = 80$), we assessed the power to correctly infer diversity dependence as the proportion of clades where the diversity-dependent model was preferred to the simpler diversity-independent model on the basis of a likelihood ratio test with a single degree of freedom and significance level of $\alpha = 0{\cdot}05$. For diversification scenarios where diversity dependence was not operating ($K = \infty$), we assessed type I error rate as the proportion of clades where the diversity-dependent model was incorrectly identified as the best fitting model. We applied this test only to simulations with $\lambda_0 = 0{\cdot}5$ and crown age = 5, because larger values often generated extremely large trees, for which likelihood optimization proved unfeasible. Finally, for both scenarios, we assessed the bias of the parameter estimates as the tendency to overestimate or underestimate the true value, and the precision as the spread of the estimated values.

When conditioning on tree size alone, we would ideally simulate clades of a fixed size. However, simulations for fixed tree size are only feasible when the phylogeny can be simulated backwards, or branching times can be sampled directly from a known probability distribution or indirectly from simulations. Simulating backwards is impossible for the diversity-dependent model, because knowledge of the diversity at any point in time is needed to calculate the diversification rates. Similarly, while sampling branching times directly is possible for various models (Stadler 2011; Höhna 2013), it is not feasible for the diversity-dependent model (Lambert & Stadler 2013). Indirect sampling, using simulation approach of Hartmann, Wong & Stadler (2010), which lets the system run to extinction, subsequently samples points at which the number of species equals the predefined size and then reconstructs the phylogeny from these points, is not feasible because when $\lambda_0 > \mu$, the diversity-dependent process is extremely unlikely to become extinct once it has reached equilibrium, and hence simulations will take a very long time. Interestingly, this problem also underlies the complexity of the computation of the likelihood conditional on fixed tree size through numerical integration: when diversity dependence is substantial, there is a non-negligible probability that the stem age of a phylogeny is in the very remote past. Given these difficulties, all our simulations assume a fixed crown age (we consider three values: 5, 10 and 15). Although we acknowledge that conditioning on tree size alone will violate this assumption, Stadler (2013) found that for the constant-rate birth–death model, results are robust towards such violations and we expect the same robustness to apply here.

Results

PARAMETER ESTIMATE ACCURACY AND PRECISION

When clades are conditioned on crown age, the parameters estimated under the diversity-dependent model, and in particular K, are subject to large biases (i.e. the average estimate across many simulations is different from the value used to generate the simulated data) and therefore cannot be reliably inferred (Figs 1, S4–S6, Supporting information). The estimate of K tends to be biased downwards towards the observed clade size, resulting in an overestimation of the true degree of clade saturation (N/K). There is also a tendency for λ_0 and μ to be overestimated. The extent of these biases is elevated when extinction is high and clades are young, corresponding to conditions in which the clades are far from equilibrium. The biases do not decrease substantially for larger trees (Figs S11–S14, Supporting information). We note that this does not conflict with asymptotic maximum-likelihood theory because using an increasingly larger number of trees to simultaneously estimate the parameters would cause the bias to decrease towards zero. Conditioning on tree size (N) also yields biased estimates, but with the bias in the opposite direction, with K overestimated, and λ_0 and μ underestimated (Figs 1, S4–S6, Supporting information). In contrast to conditioning on either crown age or N, we find that conditioning on crown age and N combined yields unbiased parameter estimates (Figs 1, S4–S6, Supporting information). However, this dual conditioning is accompanied by a severe loss of information and we find that parameters cannot be estimated precisely (i.e. we observe large variation in parameter estimates). In other words, this conditioning improves accuracy at the expense of precision. A substantial number of simulations show very unrealistic combinations of parameter estimates, with λ_0 and μ almost equal to one another, and K well below the actual tree size (Figs S11–S14, Supporting information). This means that K' is high and there is hardly any diversity dependence. We consider these parameter set artefacts of maximum likelihood.

MODEL SELECTION RELIABILITY

The χ^2 likelihood ratio test does not yield the correct type I errors (Fig. 2). That is, for a significance level of $\alpha = 0{\cdot}05$, the type I error should be, by definition, equal to 0·05. The number of times the diversity-dependent model is incorrectly selected using the likelihood ratio test is, however, much larger than α under all three conditionings, with conditioning on crown age leading to most elevated type I errors. For this reason, the power to infer diversity dependence when it is actually operating is also not represented correctly by the χ^2 likelihood ratio test. Nevertheless, differences in power for the same conditioning will still be correctly portrayed: power generally decreases as

extinction rate μ increases (Fig. 2) and as the initial speciation rate λ_0 decreases (Figs 2, S1–S3, Supporting information).

Using the likelihood ratio that corresponds with a significance level of $\alpha = 0{\cdot}05$ for the diversity-independent simulations to measure power in the diversity-dependent simulations, we find that the behaviour of the power as a function of extinction rate is similar across different parameter settings (black dots in power plots of Figs 2, S1–S3, Supporting information).

The mismatch between our type I error rate and the significance level used (Figs 2, S1–S3, Supporting information) is due to the fact that the conditions justifying the use of the χ^2 likelihood ratio test are not satisfied (Wilks 1938; Tekle, Gudicha, & Vermunt 2016; Gudicha *et al*. 2016). For proper model selection, we recommend the parametric bootstrapping approach suggested by Tekle, Gudicha, & Vermunt (2016); Gudicha *et al*. (2016). The steps involved in this bootstrap approach, applied to our case of inferring diversity dependence, and incorporating a small correction suggested by North, Curtis & Sham (2002), are as follows:

1 Estimate the maximum-likelihood (ML) parameters under the constant-rate (CR) model and the diversity-dependent (DD) model and calculate the likelihood ratio (LR). We term this LR_O.

2 Simulate X_{CR} times under CR with the ML parameters of the observed data estimated under the CR model in step 1.

3 Estimate the ML parameters for these X_{CR} simulated CR data sets under both CR and DD, and calculate the LR for each simulated data set.

4 Compare the observed LR_O with the distribution of LRs from the simulations. If the number of simulations with a larger LR than LR_O is denoted by R_{CR}, then the *p*-value of the test is $(R_{CR} + 1)/(X_{CR} + 1)$.

5 Use a significance level of α (e.g. 0·05) to accept or reject the CR model (type I error) and record the LR associated with this alpha, call this LR_α. Note that this is the $100*(1 - \alpha)$th percentile of the LRs.

6 Simulate X_{DD} times under DD with the ML parameters of the observed data estimated under the DD model in step 1.

7 Estimate parameters for these X_{DD} simulations under both CR and DD using ML and calculate the LR for each simulated data set.

8 If the number of the X_{DD} simulations where the LR exceeds LR_α is denoted by R_{DD}, then the power of the test is given by $R_{DD}/(X_{DD} + 1)$.

This analysis is computationally time-consuming and therefore precludes its application to our simulated trees, because

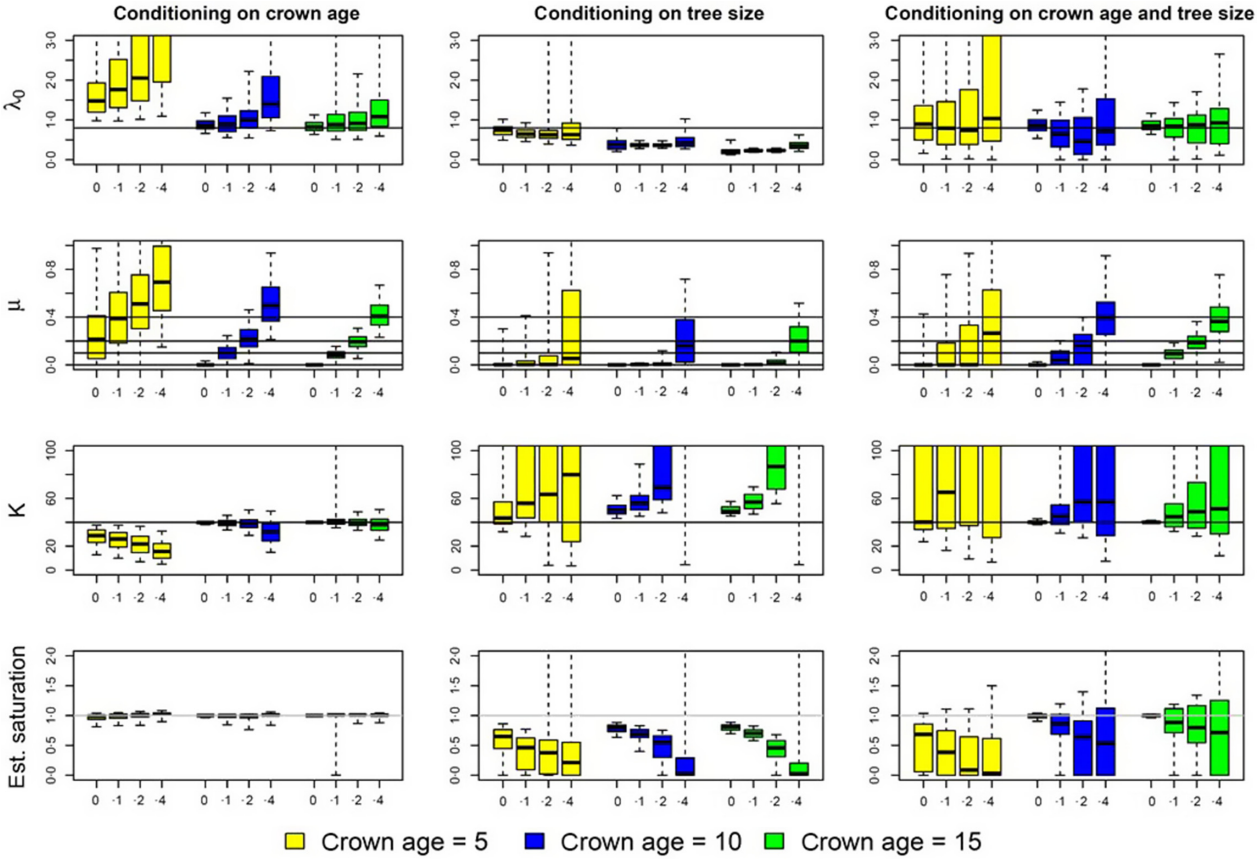

Fig. 1. Bias and precision of parameter (λ_0, μ, *K*, *N*/*K*) estimates obtained from diversity-dependent diversification models across 1000 simulations for a range of diversification scenarios (crown age: 5, 10, 15; extinction rate μ: 0, 0·1, 0·2, 0·4; intrinsic speciation rate $\lambda_0 = 0{\cdot}8$ and clade-level carrying capacity $K = 40$), for three forms of conditioning. Thin horizontal black lines correspond to the true values used to generate the simulated data. The grey line in the bottom row corresponds to the value expected in equilibrium. In the box plots, thick solid lines, boxes and whiskers correspond to the 50%, 75% and 95% percentiles, respectively.

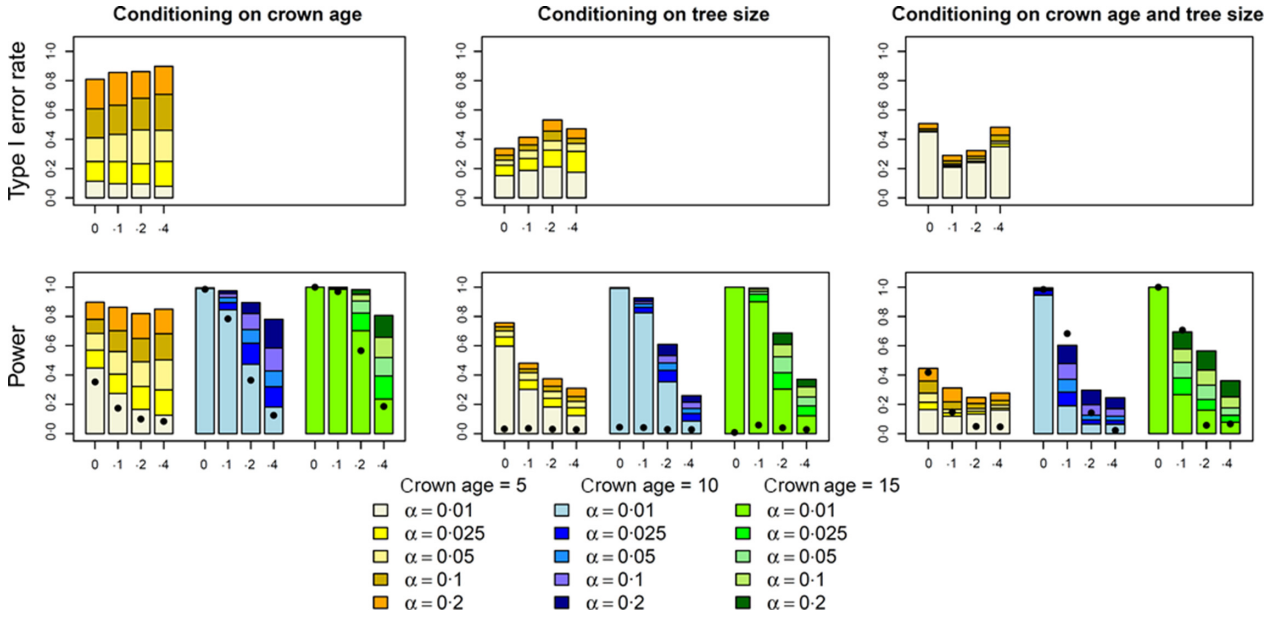

Fig. 2. Type I error rate (first row) and power (second row), based on a χ^2 likelihood ratio test, of diversity-dependent diversification models across 1000 simulations with a diversity-dependent model, for a range of diversification scenarios (crown age: 5, 10, 15; extinction rate μ: 0, 0·1, 0·2, 0·4; intrinsic speciation rate $\lambda_0 = 0{\cdot}8$ ($\lambda_0 = 0{\cdot}5$ for type I error plot) and clade-level carrying capacity $K = 40$), for three forms of conditioning, and for four significance levels α (0·01, 0·025, 0·05, 0·1, 0·2). The bars are cumulative; for example, the type I error rate for the leftmost bar (crown age = 5, μ = 0) for α = 0·2 is the sum of all stacks of the bar, while for α = 0·05, it is the sum of the three lowest stacks. Note the discrepancy between type I error rates and significance levels. The dots in the power plots indicate the power if the likelihood ratio corresponding to the true type I error rate is used in model selection.

we would have had to apply it over 100 000 times (36 diversification scenarios each applied to 1000 trees and three different conditionings). However, for analysis of data sets typically used in empirical studies, this approach is practical and realistic. For illustration purposes, we applied the approach to the *Dendroica* phylogeny, for which strong evidence for diversity dependence was previously found (Rabosky & Lovette 2008a; Etienne *et al.* 2012), and which we confirm here (Fig. 3). The *P*-value (area to the right of the black arrow in Fig. 3) is very small, <0·001 (the resolution of the bootstrap), and the power of the test (area to the right of the blue arrow) is very high: 0·996. It is also clear that the LR distribution under CR deviates from a χ^2 distribution with one degree of freedom (which is a monotonically decreasing function). The ML parameter estimates under the three conditionings are very different, with conditioning on tree size and dual conditioning yielding unrealistic values (Table 1). As argued above, we consider this an artefact of the maximum-likelihood method. There is a local likelihood optimum for both dual conditioning and conditioning on tree size that has parameter values similar to those for conditioning on crown age (which appear to be relatively unbiased for this data set, Fig. S15, Supporting information).

Discussion

PARAMETER ESTIMATE ACCURACY AND PRECISION

Our simulation study demonstrates that existing methods for inferring diversity-dependent diversification dynamics are subject to several major biases. The direction and magnitude of these biases depend heavily on both the true history of species diversification and the form of conditioning used. Conditioning on the observed crown age is the most widely used approach in phylogenetic analyses, but we find that this is often associated with strong biases in the estimated rates of speciation and clade equilibrium richness. In particular, estimates of carrying capacity tend to be biased downwards leading to the mistaken inference that clades are at or near saturation when in fact richness is still increasing. These biases are greatest when clades are young or subject to high rates of extinction. The estimates for the *Dendroica* clade seem unbiased; hence, this may be interpreted as true saturation.

An alternative to conditioning on crown age is to condition on the observed tree size. However, we find that parameters estimated using this approach are also subject to major biases, but operating in the opposite direction to when conditioning on crown age. In particular, we find that estimates of clade equilibrium richness are biased upwards leading to the mistaken inference that richness is still increasing when in fact clades are at or near saturation.

Given the biased nature of parameter estimates obtained when conditioning on either crown age or tree size, we also tested the effects of conditioning on both of these states. Our results show that this form of dual conditioning leads to less biased but very imprecise parameter estimates. For young clades or those diversifying under moderate rates of extinction, this uncertainty is so large that the estimated parameters cannot be meaningfully interpreted. These results extend previous findings for the diversity-independent model (Stadler 2013) and show that published likelihood-based inferences regarding

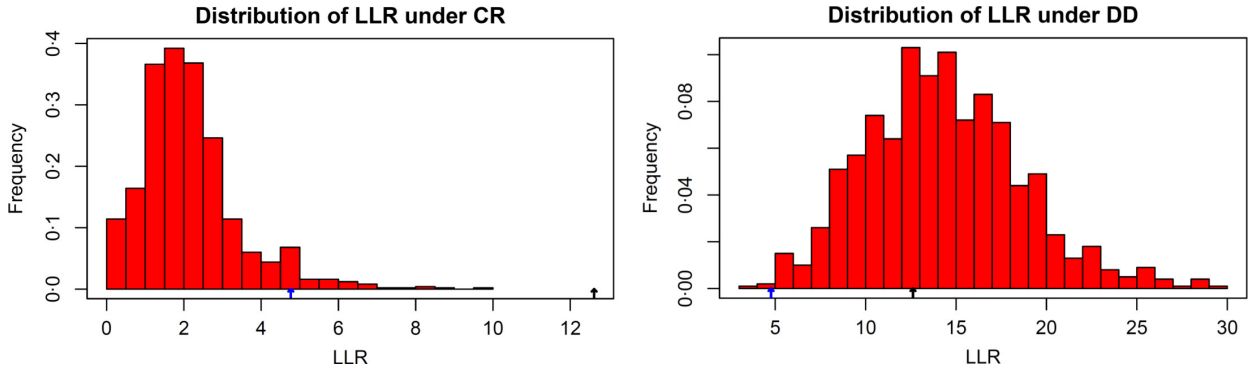

Fig. 3. Bootstrap likelihood ratio test for *Dendroica*. Left panel: the distribution of logarithms of the likelihood ratio of diversity-dependent and constant rate birth–death model for data generated under the constant-rate birth–death model. Right panel: same as top panel but for data generated under the diversity-dependent birth–death model. The black arrow shows the value of the logarithm of the likelihood ratio for the real data, while the blue arrow shows the logarithm of the likelihood ratio for a significance level of $\alpha = 0{\cdot}05$. The analysis used conditioning on crown age.

Table 1. Maximum-likelihood estimates of the DD model parameters for the *Dendroica* clade under three different conditionings

Conditioning	λ_0	μ	K
Crown age	3·04	0·17	24·59
Tree size	0·45	0·33	6·09
Crown age and tree size	0·00287	0·00279	0·656

the equilibrium richness of a clade or degree of saturation should be treated with extreme caution.

Uncertainty in parameter estimates is highest among young clades that are typically small and by themselves contain little information. This does not seem to be an effect of tree size only (Figs S11–S14, Supporting information). We caution that excluding these clades from analyses may lead to biased inferences regarding typical macroevolutionary dynamics (Ricklefs 2007; Phillimore & Price 2008). A possible approach to remove this bias is to condition on tree size having a certain minimum value, but the statistical properties of this conditioning will need to be explored.

MODEL SELECTION RELIABILITY

We found that the χ^2 likelihood ratio test to compare nested models, which is commonly used and which we employed here as well, is not valid for inferring the operation of diversity dependence, because the observed type I error does not correspond with the significance level used in the test. Using AIC as an alternative model selection tool does not resolve this, because model selection based on AIC is identical to the χ^2 likelihood ratio test in nested models: a log likelihood difference of 2 units ($\Delta LL = 2$) between two models with one parameter difference ($\Delta n = 1$) translates to an AIC difference of 2 units ($\Delta AIC = -2\ \Delta LL + 2\ \Delta n = -2$) and a *P*-value of approximately 0·05 in the χ^2 likelihood ratio test. The mismatch between type I error and significance level cannot be easily remedied by applying corrected AIC, AICc. First, the mismatch between assumed and observed type I error arises because one model is a boundary case of the other ($K = \infty$) rather than because of small sample size. Secondly, sample size is not well defined for phylogenetic trees. In particular, the likelihood of the phylogeny under a diversity-dependent model is not simply a product of likelihoods for each branching point, as is the case for the diversity-independent birth–death model after conditioning on tree size (N) (Maddison, Midford & Otto 2007; Lambert & Stadler 2013). Fortunately, proper likelihood ratio testing is still possible based on a parametric bootstrapping approach, as we outlined above. Although it is computationally costly, we anticipate that these costs are small compared to the amount of time and effort needed to collect data and build phylogenies.

The failure of the χ^2 likelihood ratio test is probably due to the fact that conditions leading to the χ^2 distribution are not met when the simpler model fixes a parameter of the more general model at a boundary of the parameter domain (Wilks 1938; Tekle, Gudicha, & Vermunt 2016; Gudicha *et al.* 2016). Here, this applies to K which is infinite for the constant rate birth–death model. However, the result may hold more generally, for example in comparisons of the constant-rate birth–death model with the pure-birth (Yule) model, or the protracted speciation model (Etienne & Rosindell 2012) to the constant rate birth–death model. We therefore recommend that our bootstrap likelihood ratio test will also be applied in these cases.

When measuring power in the diversity-dependent simulations using the likelihood ratio that corresponds to a significance level of $\alpha = 0{\cdot}05$ for the diversity-independent simulations, we found that the behaviour of the power as a function of the extinction rate is similar across different parameter settings. There may be one caveat: the parameter set on which the critical likelihood is based necessarily differs from the parameter set for which the power is computed, and these parameter sets bear no specific relationship to one another. For example, in Fig. 2, we used the critical likelihood ratio obtained from diversity-independent simulations with $\lambda_0 = 0{\cdot}5$ and $K = \infty$ for power tests of simulations with $\lambda_0 = 0{\cdot}8$ and $K = 40$, and alternative changes are used in Figs S1–S3 (Supporting information). In contrast, in our bootstrap likelihood

test, the two simulation parameter sets are related because they are the ML parameter sets for one and the same real data set. Yet, because this difference is subtle, we believe that the similarity in the behaviour of the power as a function of extinction rate is robust.

Interpreted pessimistically, our analyses suggest that under currently available forms of conditioning, parameters estimated from diversity-dependent models are either biased or subject to such high uncertainty as to be of little practical use for parameter estimation. Therefore, the use of these models should be restricted to hypothesis testing, where our recommended bootstrap procedure will identify the accuracy of the inference. Optimistically, however, the parametric bootstrap can help identify the bias and uncertainty and it is possible that new forms of conditioning will be developed that partially or even completely ameliorate these problems.

Our results provide several guidelines when attempting to infer diversity-dependent dynamics. First, if the aim of the study was hypothesis testing and to ascertain whether the dynamics of diversification have been subject to diversity dependence, then it does not seem to matter much which conditioning is used, but in all cases power may be low. Secondly, parameter values obtained from diversity-dependent models should be interpreted extremely cautiously. In the case of single conditionings, parameter estimates are strongly biased but in opposite directions, while in the case of dual conditioning, estimates are subject to extremely high uncertainty. The magnitude of this uncertainty may be greater than the bias in parameters under single conditioning. To gain a quick overview of this uncertainty, one can compare parameter estimates for various initial values and across the different forms of conditioning; for a more thorough overview, we recommend plotting the likelihood surface (around the optimum). Alternatively, one can use parametric bootstrapping to obtain parameter uncertainty estimates (Fig. S15, Supporting information; e.g. Etienne, Morlon & Lambert 2014 for the protracted speciation model). The proposed bootstrap likelihood ratio test suggested above can provide the necessary uncertainty estimates, and thus these come at no extra cost. We note that this parametric bootstrap approach is limited to conditioning on clade size only because simulations under the other two conditionings are practically impossible.

Any meta-analysis based on estimated parameters (e.g. examining how carrying capacity K or initial speciation rate varies across clades as a function of a covariate such as latitude) should explicitly account for parameter uncertainty when estimating the effect sizes and significance of these trends. The modelling framework to do this already exists, as it is possible to incorporate the squared standard error of parameter estimates into a phylogenetic meta-analysis (Ives, Midford & Garland 2007; Hadfield & Nakagawa 2010).

In this paper, we have focused on the question of whether the presence or absence of diversity dependence can be reliably detected by comparing a diversity-dependent model with its diversity-independent limit. We have not considered the question of whether detection of diversity dependence truly indicates that diversity dependence is operating, or whether other mechanisms (e.g. time-dependent speciation rates) can be mistaken for diversity dependence or vice versa. Future research dealing with this important issue will have to resolve the complication that comparison of non-nested models may strongly depend on the mathematical function used to implement the mechanism.

Whether species richness exists at some form of steady state or increases unbounded is one of the most fundamental questions we can ask about biodiversity. Along with the fossil record, reconstructed phylogenies are among the few sources of information that biologists have found to address this question. We therefore encourage further research in the development of methods that fully utilize the information encoded within phylogenies to provide reliable and robust inferences on the dynamics of species diversification.

Acknowledgements

We thank T. Stadler, D. Rabosky and F. Bokma for helpful comments. RSE and ALP thank the Netherlands Organisation for Scientific Research (NWO) for financial support through VICI and VENI grants, and ABP is funded by a NERC fellowship Ne/I020598/1.

Data accessibility

Code and results of simulations and maximum-likelihood analysis. Dryad entry doi:10.5061/dryad.tg37f (Etienne 2016).

References

Alfaro, M.E., Santini, F., Brock, C., Alamillo, H., Dornburg, A., Rabosky, D.L., Carnevale, G., Harmon, L.J. & Hillis, D.M. (2009) Nine exceptional radiations plus high turnover explain species diversity in jawed vertebrates. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 13410–13414.

Bokma, F. (2009) Problems detecting density-dependent diversification on phylogenies. *Proceedings of the Royal Society of London B: Biological Sciences*, **276**, 993–994.

Etienne, R.S. & Haegeman, B. (2012) A conceptual and statistical framework for adaptive radiations with a key role for diversity dependence. *American Naturalist*, **180**, 75–89.

Etienne, R.S., Morlon, H. & Lambert, A. (2014) Estimating the duration of speciation from phylogenies. *Evolution*, **68**, 2430–2440.

Etienne, R.S. & Rosindell, J. (2012) Prolonging the past counteracts the pull of the present: protracted speciation can explain observed slowdowns in diversification. *Systematic Biology*, **61**, 204–213.

Etienne, R.S., Haegeman, B., Stadler, T., Aze, T., Pearson, P.N., Purvis, A. & Phillimore, A.B. (2012) Diversity-dependence brings molecular phylogenies closer to agreement with the fossil record. *Proceedings of the Royal Society of London B: Biological Sciences*, **279**, 1300–1309.

Etienne, R.S. (2016) How reliably can we infer diversity-dependent diversification from phylogenies? *Dryad Digital Repository*. doi: 10.5061/dryad.tg37f

Gudicha, G.W., Schmittmann, V.D., Tekle, F.B. & Vermunt, J.K. (2016) Power analysis for the likelihood-ratio test in latent Markov models: short-cutting the bootstrap p-value based method. *Multivariate Behavioral Research*, In press.

Hadfield, J.D. & Nakagawa, S. (2010) General quantitative genetic methods for comparative biology: phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *Journal of Evolutionary Biology*, **23**, 494–508.

Harmon, L.J. & Harrison, S. (2015) Species diversity is dynamic and unbounded at local and continental scales. *American Naturalist*, **185**, 584–593.

Hartmann, K., Wong, D. & Stadler, T. (2010) Sampling trees from evolutionary models. *Systematic Biology*, **59**, 465–476.

Höhna, S. (2013) Fast simulation of reconstructed phylogenies under global time-dependent birth–death processes. *Bioinformatics*, **29**, 1367–1374.

Ives, A.R., Midford, P.E. & Garland, T. (2007) Within-species variation and measurement error in phylogenetic comparative methods. *Systematic Biology*, **56**, 252–270.
Jønsson, K.A., Fabre, P.-H., Fritz, S.A., Etienne, R.S., Ricklefs, R.E., Jørgensen, T.B. *et al.* (2012) Ecological and evolutionary determinants for the adaptive radiation of the Madagascan vangas. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 6620–6625.
Lambert, A. & Stadler, T. (2013) Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. *Theoretical Population Biology*, **90**, 113–128.
Maddison, W.P., Midford, P.E. & Otto, S.P. (2007) Estimating a binary character's effect on speciation and extinction. *Systematic Biology*, **56**, 701–710.
McInerny, G.J. & Etienne, R.S. (2012a) Ditch the niche - is the niche a useful concept in ecology or species distribution modelling? *Journal of Biogeography*, **39**, 2096–2102.
McInerny, G.J. & Etienne, R.S. (2012b) Stitch the niche - a practical philosophy and visual schematic for the niche concept. *Journal of Biogeography*, **39**, 2103–2111.
McInerny, G.J. & Etienne, R.S. (2012c) Pitch the niche - taking responsibility for the concepts we use in ecology and species distribution modelling. *Journal of Biogeography*, **39**, 2112–2118.
Moen, D. & Morlon, H. (2014) Why does diversification slow down? *Trends in Ecology & Evolution*, **29**, 190–197.
Morlon, H. (2014) Phylogenetic approaches for studying diversification. *Ecology Letters*, **17**, 508–525.
Nee, S., May, R.M. & Harvey, P.H. (1994) The reconstructed evolutionary process. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **344**, 305–311.
North, B.V., Curtis, D. & Sham, P.C. (2002) A note on the calculation of empirical p values from Monte Carlo procedures. *American Journal of Human Genetics*, **71**, 439–441.
Phillimore, A.B. & Price, T.D. (2008) Density-dependent cladogenesis in birds. *PLoS Biology*, **6**, e71.
Pigot, A.L., Phillimore, A.B., Owens, I. & Orme, C.D.L. (2010) The shape and temporal dynamics of phylogenetic trees arising from geographic speciation. *Systematic Biology*, **59**, 660–673.
Pybus, O.G. & Harvey, P.H. (2000) Testing macro-evolutionary models using incomplete molecular phylogenies. *Proceedings of the Royal Society of London B: Biological Sciences*, **267**, 2267–2272.
Pyron, R.A. & Wiens, J.J. (2013) Large-scale phylogenetic analyses reveal the causes of high tropical amphibian diversity. *Proceedings of the Royal Society of London B: Biological Sciences*, **280**, 20131622.
Rabosky, D.L. (2009) Ecological limits and diversification rate: alternative paradigms to explain the variation in species richness among clades and regions. *Ecology Letters*, **12**, 735–743.
Rabosky, D.L. (2013) Diversity-dependence, ecological speciation, and the role of competition in macroevolution. *Annual Review of Ecology Evolution and Systematics*, **44**, 481–502.
Rabosky, D.L. & Lovette, I.J. (2008a) Density-dependent diversification in North American wood warblers. *Proceedings of the Royal Society of London B: Biological Sciences*, **275**, 2363–2371.
Rabosky, D.L. & Lovette, I.J. (2008b) Explosive evolutionary radiations: decreasing speciation or increasing extinction through time? *Evolution*, **62**, 1866–1875.
Ricklefs, R.E. (2007) Estimating diversification rates from phylogenetic information. *Trends in Ecology & Evolution*, **22**, 601–610.
Rundell, R.J. & Price, T.D. (2009) Adaptive radiation, nonadaptive radiation, ecological speciation and nonecological speciation. *Trends in Ecology & Evolution*, **24**, 394–399.
Stadler, T. (2011) Simulating trees with a fixed number of extant species. *Systematic Biology*, **60**, 676–684.
Stadler, T. (2013) How can we improve accuracy of macroevolutionary rate estimates? *Systematic Biology*, **62**, 321–329.
Tekle, F.B., Gudicha, G.W. & Vermunt, J.K. (2016) Power analysis forthe bootstrap likelihood ratio test in latent class models. *Advances in Classification and Data Analysis*, In press.
Valente, L.M., Phillimore, A.B. & Etienne, R.S. (2015) Equilibrium and non-equilibrium dynamics simultaneously operate in the Galápagos islands. *Ecology Letters*, **18**, 844–852.
Weir, J.T. (2006) Divergent timing and patterns of species accumulation in lowland and highland neotropical birds. *Evolution*, **60**, 842–855.
Wilks, S.S. (1938) The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Annals of Mathematical Statistics*, **9**, 60–62.

Received 12 October 2015; accepted 9 March 2016
Handling Editor: Kate Jones

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Fig. S1. Power as in Fig. 2, but for intrinsic speciation rate $\lambda_0 = 0.5$ and clade-level carrying capacity $K = 40$.

Fig. S2. Power as in Fig. 2, but for intrinsic speciation rate $\lambda_0 = 0.5$ and clade-level carrying capacity $K = 80$.

Fig. S3. Power as in Fig. 2, but for intrinsic speciation rate $\lambda_0 = 0.8$ and clade-level carrying capacity $K = 80$.

Fig. S4. Parameter estimates as in Fig. 1, but for intrinsic speciation rate $\lambda_0 = 0.5$ and clade-level carrying capacity $K = 40$.

Fig. S5. Parameter estimates as in Fig. 1, but for intrinsic speciation rate $\lambda_0 = 0.5$ and clade-level carrying capacity $K = 80$.

Fig. S6. Parameter estimates as in Fig. 1, but for intrinsic speciation rate $\lambda_0 = 0.8$ and clade-level carrying capacity $K = 80$.

Fig. S7. Supplementary statistics of the simulated data for three forms of conditioning (columns) across different diversification scenarios (1000 simulations).

Fig. S8. Supplementary statistics as in Fig. S7 but for intrinsic speciation rate $\lambda_0 = 0.5$ and clade-level carrying capacity $K = 40$.

Fig. S9. Supplementary statistics as in Fig. S7, but for intrinsic speciation rate $\lambda_0 = 0.5$ and clade-level carrying capacity $K = 80$.

Fig. S10. Supplementary statistics as in Fig. S7, but for intrinsic speciation rate $\lambda_0 = 0.8$ and clade-level carrying capacity $K = 80$.

Fig. S11. Estimates of K as a function of the size of the simulated tree for different diversification scenarios and forms of conditioning (1000 simulations).

Fig. S12. Relationship of the estimated K vs. tree size as in Fig. S11, but for intrinsic speciation rate $\lambda_0 = 0.5$ and clade-level carrying capacity $K = 40$.

Fig. S13. Relationship of the estimated K vs. tree size as in Fig. S11, but for intrinsic speciation rate $\lambda_0 = 0.5$ and clade-level carrying capacity $K = 80$.

Fig. S14. Relationship of the estimated K vs. tree size as in Fig. S11, but for intrinsic speciation rate $\lambda_0 = 0.8$ and clade-level carrying capacity $K = 80$.

Fig. S15. Bootstrap distribution of the parameter values for the *Dendroica* clade.